Journal of Structural and Functional Genomics **14:** 53–66, 2002.

Crystal structure of a conserved hypothetical protein from *Escherichia coli*

Dong Hae Shin[1], Hisao Yokota[2], Rosalind Kim[2] & Sung-Hou Kim[1,2,*]
[1]*Department of Chemistry, University of California, Berkeley, California 94720-5230, USA;* [2]*Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA;* ∗*Author for correspondence (e-mail: SHKim@cchem.berkeley.edu; fax: 1-510-486-5272)*

Received 7 September 2001; Accepted in revised form 23 November 2001

Key words: hypothetical protein, structural genomics, crystal structure, new fold, conserved cysteine

Abstract

The crystal structure of a conserved hypothetical protein from *Escherichia coli* has been determined using X-ray crystallography. The protein belongs to the Cluster of Orthologous Group COG1553 (National Center for Biotechnology Information database, NLM, NIH), for which there was no structural information available until now. Structural homology search with DALI algorism indicated that this protein has a new fold with no obvious similarity to those of other proteins with known three-dimensional structures. The protein quaternary structure consists of a dimer of trimers, which makes a characteristic cylinder shape. There is a large closed cavity with approximate dimensions of 16 Å × 16 Å × 20 Å in the center of the hexameric structure. Six putative active sites are positioned along the equatorial surface of the hexamer. There are several highly conserved residues including two possible functional cysteines in the putative active site. The possible molecular function of the protein is discussed.

Abbreviations:

Introduction

A large amount of genomic sequence information has been provided by completed and ongoing large-scale genome sequencing projects [http://www.tigr.org/tdb/mdb/mdb.html, http://www.mcs.anl.gov/home/gaasterl/genomes.html]. In many cases, the function of the encoded gene products can be deduced from comparative sequence analysis [1]. However, for a large fraction of the predicted gene products, no functions can be inferred because of the absence of reliable sequence similarity to proteins with known function [2]. Other methods using information such as phylogenetic profiles, domain fusions, and gene localization can sometimes provide information about cellular function [3] with less reliabilities.

Since the three-dimensional structure of a protein is tightly coupled to its molecular (biochemical and biophysical) function, the structure of a protein with unknown (based on sequence information alone) function may infer its molecular function. A number of recent publications demonstrated the validity of this approach [4-6], although in some cases it is more difficult to infer molecular functions when the structure has a novel, new fold [7, 8].

An open reading frame in the chaC-narL intergenic region of *Escherichia coli (E. coli)* codes for a hypothetical protein of 12.7 kDa molecular weight named ychN [9]. A PSI-BLAST search with this sequence against the database nr (all non-redundant GenBank CDS translations + PDB + SwissProt + PIR + PRF) revealed seven sequences with full-length homology with sequence identity ranging from 28 to 88% (Figure 1). This group of proteins was named the ychN family [9]. Most of the homologous sequences are annotated as hypothetical proteins. YchN belongs to a family of conserved hypothetical proteins known as COG1553 in the National Center for Biotechnology Information (NCBI) database of Clusters of Orthologous Groups [10]. COG1553 includes yheN from *E. coli*, HI0576 from *Haemophilus influenza Rd*,

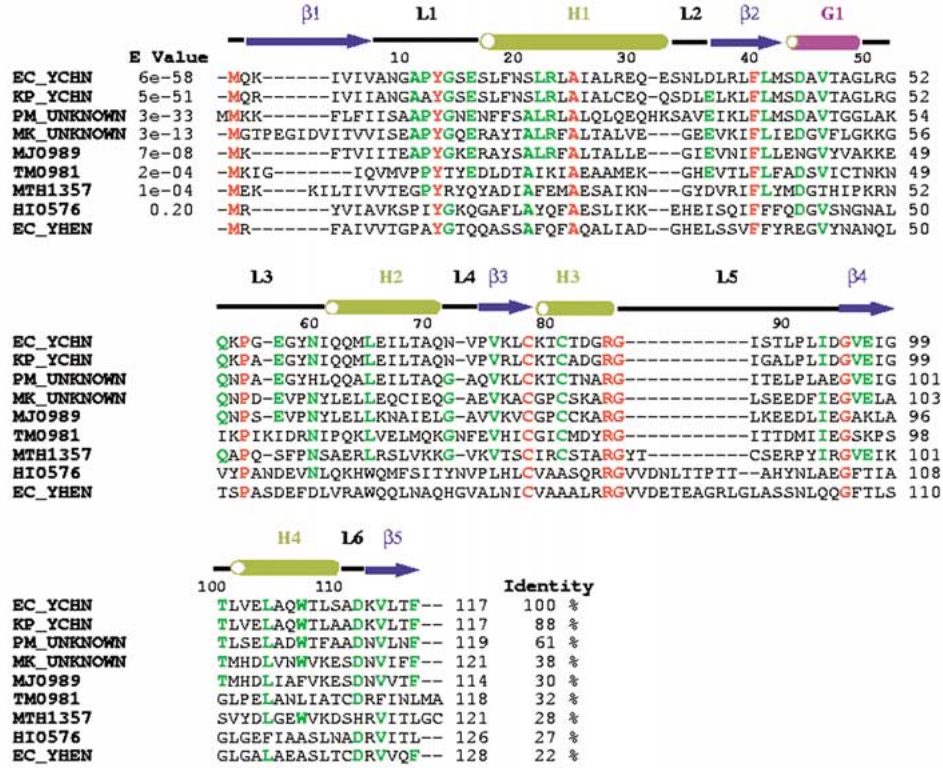

Figure 1. Sequence comparison among ychN and its homologs. YchN from *Escherichia coli* (EC_YCHN), ychN from *Klebsiella pneumoniae* (KP_YCHN), unknown protein from *Pasteurella multocida* (PM_UNKNOWN), unknown protein from *Methanopyrus kandleri* (MK_UNKNOWN), Mj0989 from *Methanococcus jannaschii* (MJ0989), TM0981 from *Thermotoga maritima* (TM0981), MTH1357 from *Methanothermobacter thermoautotrophicus* (MTH1357), HI0576 from *Haemophilus influenza Rd* (HI0576), and yheN from *E. coli* (EC_YHEN). The first seven sequences were detected with an E-value better than threshold (E = 0.01) by running PSI-BLAST. HI0576 was detected with an E-value worse than the threshold and yheN was not detected in PSI-BLAST. Invariant residues are highlighted as red characters. Green characters represent residues conserved in five or more proteins out of seven sequences having an E-value better than threshold. The '-'s represent gaps.

TM0981 from *Thermotoga maritima*, ychN from *E. coli*, MJ0989 from *Methanococcus jannaschii* (Mj), and MTH1357 from *Methanothermobacter thermoautotrophicus*. Among these, HI0576 and yheN sequences were not detectable by PSI-BLAST with an expectation value better than E = 0.01 [1]. However, all of the members are assumed to have an uncharacterized ancient conserved region (ACR; common to two or more main phylogenetic branches - archaea, bacteria, eukarya) involved in intracellular sulfur reduction [10]. In particular, HI0576 and yheN have a strong sequence homology with DsrE protein from *Chromatium vinosum* D [11]. DsrE is a member of the gene cluster *dsrABEFHCMK*, products of which conduct intracellular oxidation of stored sulfur [11]. These sequence analyses indicate a possible function for the cysteine residues in ychN protein. We have determined the three-dimensional structure of ychN by X-ray crystallography and discuss a possible molecular function of this family.

Materials and methods

Bacterial expression and protein purification

E. coli ychN protein was unintentionally co-purified during the purification of another hypothetical protein from *Methanococcus jannaschii*, Mj0608. The gene for Mj0608 was amplified from genomic *M. jannaschii* DNA, cloned into pET21a (Novagen, Madison, WI), and was confirmed by DNA sequencing. A selenomethionine derivative of the protein was expressed in a methionine auxotroph, *E. coli* strain B834 (DE3)/pSJS1244 [12, 13], and grown in M9 medium supplied with selenomethionine. In the purification process, the cell lysate was subjected to heating (80 °C

for 30 min). After heating, anion exchange (HiTrap-Q) column chromatography was performed three times. The protein was eluted in 50 mM Tris-HCl, pH 7.2, 150 mM NaCl. All purification steps include 10 mM DTT to avoid potential oxidation of selenomethionine. The combined yield of Mj0608 and ychN was typically 5 mg of pure protein/liter of culture. The SDS-PAGE showed one band around 34 kDa corresponding to the molecular weight of Mj0608 and dynamic light scattering also confirmed this by showing a monodisperse peak of a size similar to that of Mj0608. As the initial crystallization screens with 10 mg/ml of the proteins indicated insufficient protein concentration when using salt as a precipitant, the protein concentration was gradually increased to 250 mg/ml. The initial crystallization conditions were screened by the sparse matrix method using the Hampton Research kits (Laguna Niguel, California) [14] and by in-house developed screening methods (unpublished) at 22 ($\pm$0.5) °C. One microliter of 250 mg/ml of the protein mixture in 50 mM Tris $\cdot$ HCl, pH 7.2, 150 mM NaCl, was mixed with one microliter of 2.4 M ammonium phosphate, 0.1 M citric acid at pH 5.5. The hanging drop was equilibrated with 0.5 ml of 2.4 M ammonium phosphate, 0.1 M citric acid at pH 5.5. Rod shaped crystals grew in a month to approximate dimensions of 0.1 mm $\times$ 0.02 mm $\times$ 0.02 mm.

Data collection and reduction

The crystals were soaked in a drop of mother liquor with 16% glycerol (about 10 μl) for about one minute before being flash-frozen in liquid nitrogen and exposed to X-ray. X-ray diffraction data sets were collected at three wavelengths at the Macromolecular Crystallography Facility beamline 5.0.2 at the Advanced Light Source at Lawrence Berkeley National Laboratory using an Area Detector System Co. (ADSC) Quantum 4 CCD detector placed 140 mM from the sample. The oscillation range per image was 1.0° with no overlap between two contiguous images. X-ray diffraction data were processed and scaled using DENZO and SCALEPACK from the HKL program suite [15]. The synchrotron data were collected to 2.8 Å. Data statistics are summarized in Table 1a. The crystal belongs to the primitive orthorhombic space group $P2_12_12_1$, with unit-cell parameters of a = 66.21 Å, b = 80.46 Å, and c = 140.15 Å.

Structure determination and refinement

The program SOLVE [16] was used to locate the selenium sites in the crystal and to calculate initial phases. The initial multi-wavelength anomalous dispersion phases were further improved by solvent flattening and histogram matching with the DM program in the CCP4 package [17]. The map calculated by using the improved phases was not good enough to trace the backbone structure of the protein. However, the presence of a two-fold non-crystallographic symmetry (NCS) was recognized among selenium sites, and two-fold NCS density averaging was carried out using DM [17]. The improved density map revealed that there were six subunits in the asymmetric unit, and the six partial models were built using the O program [18]. Based on these, six NCS matrices were found and these matrices were refined using software in the RAVE package of real-space averaging and density-manipulation programs [19]. During backbone tracing and sequence fitting with the improved map, we realized that the electron density did not correspond to the amino acid sequence of Mj0608. The six-fold NCS averaging using DM gave a high quality electron density map. Two subunits showed three selenomethionine electron densities including the N-terminal selenomethionine residue. The other four subunits showed only two selenomethionines because of the flexibility of the N-terminal selenomethionine residues. A polyalanine model of 115 residues was built for each subunit. Based on this structure, we could get information about the relative positions of the methionine residues. The distance between two methionine residues is 39 amino acids and the other interval corresponds to 21 amino acid residues.

To identify the protein in the *E. coli* genome, the proteins satisfying the following two conditions were searched: (1) proteins containing two to four methionines, and (2) proteins containing 115 to 120 amino acid residues. This search produced 48 candidates. As we knew the relative positions of the methionines, the relative gaps between methionine residues were surveyed among the selected candidates. Only the ychN sequence satisfied both conditions, although the exact intervals between methionines are 40 and 21 residues. Sequence fitting of ychN matched the electron density map.

The program CNS was used for all refinement calculations [20]. All the reflections in the remote data set between 20.0 Å and 2.8 Å were included throughout the refinement calculations. The non-crystallographic

Table 1. Statistics of X-ray diffraction data and structure refinement.

Data set	Edge	Peak	Remote
(a) Statistics of the three wavelength MAD data sets			
Wavelength (Å)		0.97949	0.97926	0.99999
Resolution (Å)		70.1–2.8	70.1–2.8	70.1–2.8
Redundancy		3.08 (2.91)[a]	3.08 (2.93)	3.07 (2.93)
Unique reflections		35137 (1784)	35145 (1799)	35268 (1756)
Completeness (%)		97.2 (98.7)	97.2 (98.8)	97.3 (98.6)
I/σ		10.2 (3.8)	10.2 (4.0)	10.6 (3.9)
Rsym[b] (%)		10.8 (31.4)	10.5 (28.1)	10.6 (32.5)
(b) Crystal parameters and refinement statistics			
The remote data set was used for structure refinement			
Space group		$P2_12_12_1$	
Cell dimensions		a = 66.2 Å; b = 80.5 Å; c = 140.2 Å	
Volume fraction of protein		47.9%	
V_m (Å^3/Dalton)		2.45	
Total number of residues		702	
Total non-H atoms		5334	
Number of water molecules		86	
Average temperature factors			
Protein		25.3 Å^2	
Solvent		37.1 Å^2	
Resolution range of reflections used		20.0–2.8 Å	
Amplitude cutoff		No	
R factor		20.8%	
Free R factor		24.9%	
Stereochemical ideality:	bond	0.008 Å	
	angle	1.40°	
	improper	0.83°	
	dihedral	23.97°	

[a]Numbers in parenthesis refer to the highest resolution shell, which is 2.80–2.85 Å for all wavelength data. [b]$R_{\text{sym}} = \sum_{hkl} \sum_i |I_{hkl,i} - \langle I \rangle_{hkl}| / \sum |\langle I \rangle_{hkl}|$.

symmetry restraints among six subunits were applied during the refinement. Ten percent of the data were randomly chosen for free R factor cross validation. The refinement statistics are shown in Table 1b. Isotropic B-factors for individual atoms were initially fixed to 15 Å^2 and were refined in the last stages. The 2Fo–Fc and Fo–Fc maps were used for the manual rebuilding between refinement cycles and for the location of solvent molecules. When the refined B-factor of a solvent molecule exceeded 60 Å^2, it was removed. Atomic coordinates have been deposited in to the Protein Data Bank (PDB) with the access code of 1jx7.

Results

Quality of the model

All residues are well defined by the electron density for the refined models of ychN (Figure 2). The final model has been refined at 2.8 Å resolution to a crystallographic R-factor of 20.8%. The root-mean-square (r.m.s.) deviations from ideal stereochemistry are 0.008 Å for bond lengths, 1.40° for bond angles and 0.83° for improper angles. The averaged B-factors for main chain atoms and side chain atoms are 24.7 Å^2 and 25.9 Å^2, respectively. In the model of the ychN, the residues showing higher B-factors are located around the short loop between first α-helix and

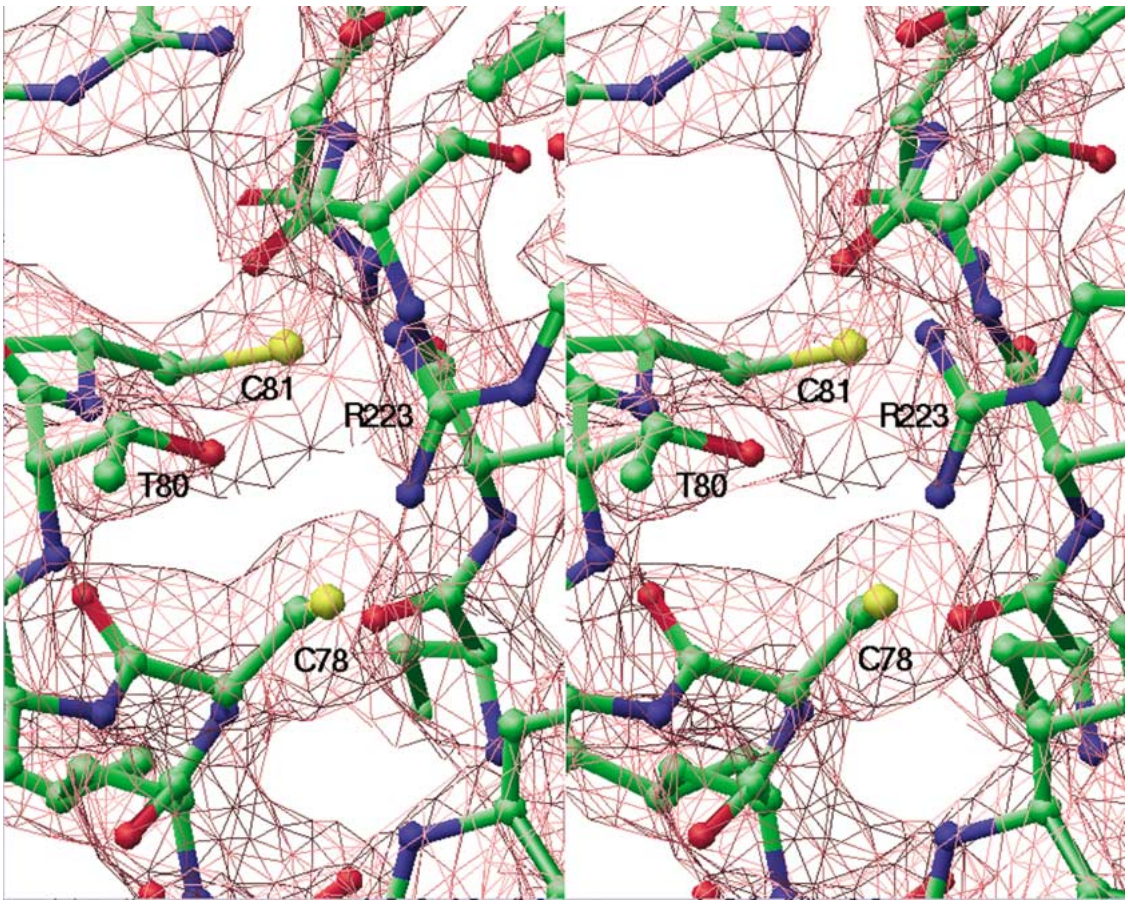

Figure 2. A stereo view of the initial 6-fold NCS averaged electron-density map countered at 1σ. The 2fo-fc map from finally refined phase was calculated using all reflection data between 20 Å and 2.8 Å. The figure was generated using the program RIBBONS [38]. The dark red represents the electron-density map. Blue represents nitrogen atoms, red for oxygen, yellow for sulfur, and green for carbon.

second β-strand (residues, 31–34; 50.7 Å^2), and in the large loop between 3_{10}-helix and second α-helix (residues 54–58; 41.4 Å^2). Table 1b summarizes the refinement statistics as well as model quality parameters. The mean positional error in atomic coordinates for refined model is estimated to be within 0.30 Å by the Luzzati plot [21]. All residues lie in the allowed region of the Ramachandranplot produced with PROCHECK [22].

Overall structure

A Cα trace of a monomer is shown in Figure 3a. The monomer has approximate dimensions of 36 Å × 30 Å × 28 Å. The central β-sheet is made of five β-strands [β1 (residues 2–8), β2 (residues 36–42), β3 (residues 74–78), β4 (residues 95–99), and β5 (residues 113–117)] (Figure 3b). The sheet is surrounded by one 3_{10}-helix [G1 (residues 43–49)] and four α-helices [H1 (residues 16–32), H2 (residues 61–71), H3 (residues 79–85), H4 (101–110)] connected by six loops [L1 (residues 9–15), L2 (residues 33–35), L3 (residues 50–60), L4 (residues 72–73), L5 (residues 86–94), and L6 (residues 111–112)].

The crystal structure shows a dimer of homotrimers consisting of six essentially identical subunits (Figure 4). The dimensions of the hexamer are 55 Å along the triad axis and 72 Å across the 3-fold axis. The homotrimer is stabilized by hydrophobic interactions between residues from amphipathic helix H1 (residues Leu24, Ala27, and Leu28) and strand β5 (residues Leu115 and Phe117) of one subunit and residues from helix H4 (Leu101, Ala105, and Leu109) of the neighboring subunit (Figure 5a). The interactions among the β-strands (β5s) of each subunit in the center of trimer also seem to help trimer stabilization (Figure 5b).

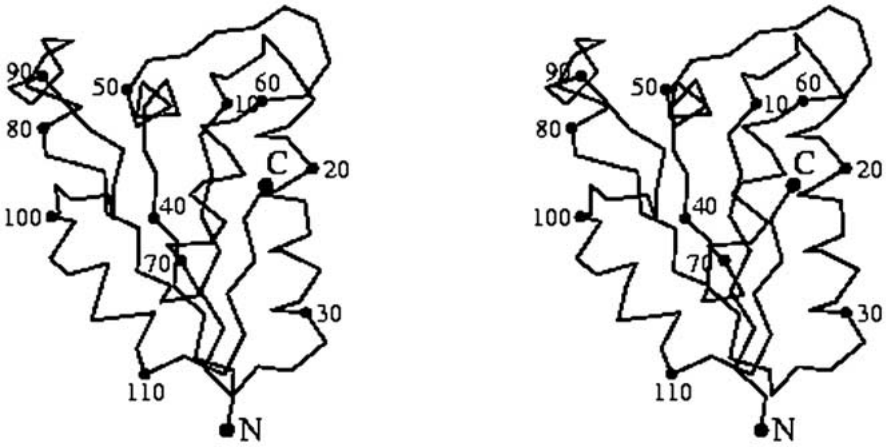

Figure 3a. A stereo drawing of a Cα trace of ychN. Every tenth residue is numbered and represented by a dot. The N and C termini are labeled. The figure was generated by MOLSCRIPT [39].

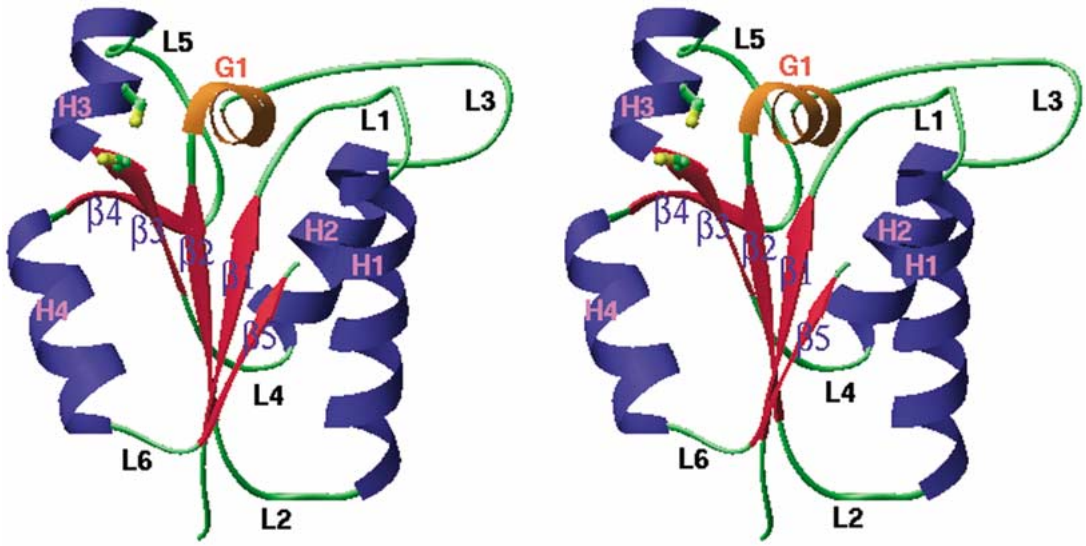

Figure 3b. A ribbon diagram of ychN. The figure was drawn in the same orientation as Figure 3a. α-helices are shown in blue, β-strands in red, and 3_{10}-helix in gold. All secondary structure elements are labeled. Two highly conserved cysteine residues are represented by a ball-and-stick model.

The trimer-trimer interaction on the equatorial interface is formed mainly by the inner L1–L1′ and outer L3–L3′ loop interactions (Figure 4). Residues Thr47, Leu50, Ile87, and Leu90 of each subunit make a hydrophobic patch with the same residues of a two-fold related subunit that seems to mediate stable trimer-trimer interaction (Figure 4 and Figure 6). The crystal structure reveals a homohexameric complex with the dihedral point group symmetry D3 (Figure 4). Two layers of three protein chains form a cylindrical arrangement with a central hole. This hole is a closed cavity in the core of the complex with approximate dimensions of 16 Å × 16 Å × 20 Å.

The structure suggests that the ychN hexamer may be a functional oligomer rather than a crystallization artifact for several reasons: (1) the interactions among monomers in the hexamer are very tight and highly symmetric; (2) there are many nonpolar interactions among monomers as commonly observed in biologically functional oligomer interfaces [23]; and (3) the surface area buried by the creation of a hexamer (Table 2) is consistent with those from other oligomeric

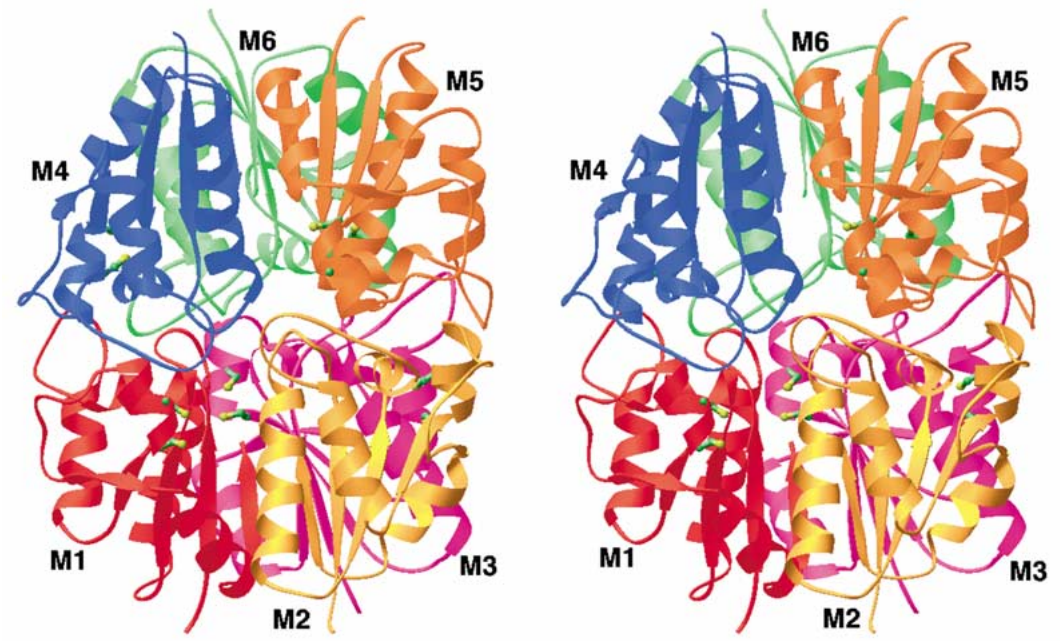

Figure 4. A ribbon diagram of the ychN hexamer. Each of the six subunits is colored differently and labeled. Two highly conserved cysteine residues are represented by a ball-and-stick model.

Table 2. Buried accessible surface area calculation.

	Total surface area (Å^2)	Area buried per monomer (Å^2)
Monomer	5026.7	
Dimer	9156.2	448.8
Trimer	12291.2	929.7
Hexamer	21320.0	1473.4

proteins and is larger than the area buried by typical crystal contacts [23].

The side chain atoms of the following amino acid residues form the inner surface of the cavity: Val7, Ala8, Asn9, Gly10, Gly14, Ser15, Glu16, Ser17, Met42, Ser43, Asp44, Thr116, and Phe117. Interestingly, Glu16 and Asp44 are located on the borderline of a possible active site (discussed below), and are highly conserved by sequence alignment (Figure 1). Therefore, the inside surface wall is negatively charged (Figure 6). There are many water molecules around protein residues lining the cavity.

There are eight invariant amino acids beside the N-terminal methionine as shown by sequence alignment (Figure 1). All amino acids, except Cys78, appear to be essential for structural reasons. Cys78 is located in a possible active site and may be an active residue as discussed below.

Discussion

Folding topology

The ychN protein has a new fold not found in PDB. It is composed of β1, α1, β2, G1, α2, α3, β3, β4, α4 and β5 (Figure 3b). The central β-sheet of ychN has a topology of $-4x$, $+1x$, $+1x$, $+1x$ [24] or topology $5_\uparrow 1_\uparrow 2_\uparrow 3_\uparrow 4_\uparrow$ [25] (Figure 7a). This β-sheet may be considered as an extension of a four-stranded β-sheet with motif $4_\uparrow 1_\uparrow 2_\uparrow 3_\uparrow$ in a classic Rossmann fold. However, a five-stranded β-sheet with topology $5_\uparrow 1_\uparrow 2_\uparrow 3_\uparrow 4_\uparrow$ has not previously been observed [25].

The structural classification databases CATH [26], DALI/FSSP [27] and SCOP [28] were utilized for comparative analysis of the ychN structure. The fold of ychN resembles many others with the three layer β-sandwich form, a fold in a α+β class. We searched for its structural homologs in the PDB with the program DALI. As expected from the low molecular weight and simple secondary structure of ychN, the DALI search revealed 292 candidates that show a z score above 2.0 (112 for z above 3.0, 30 for z above

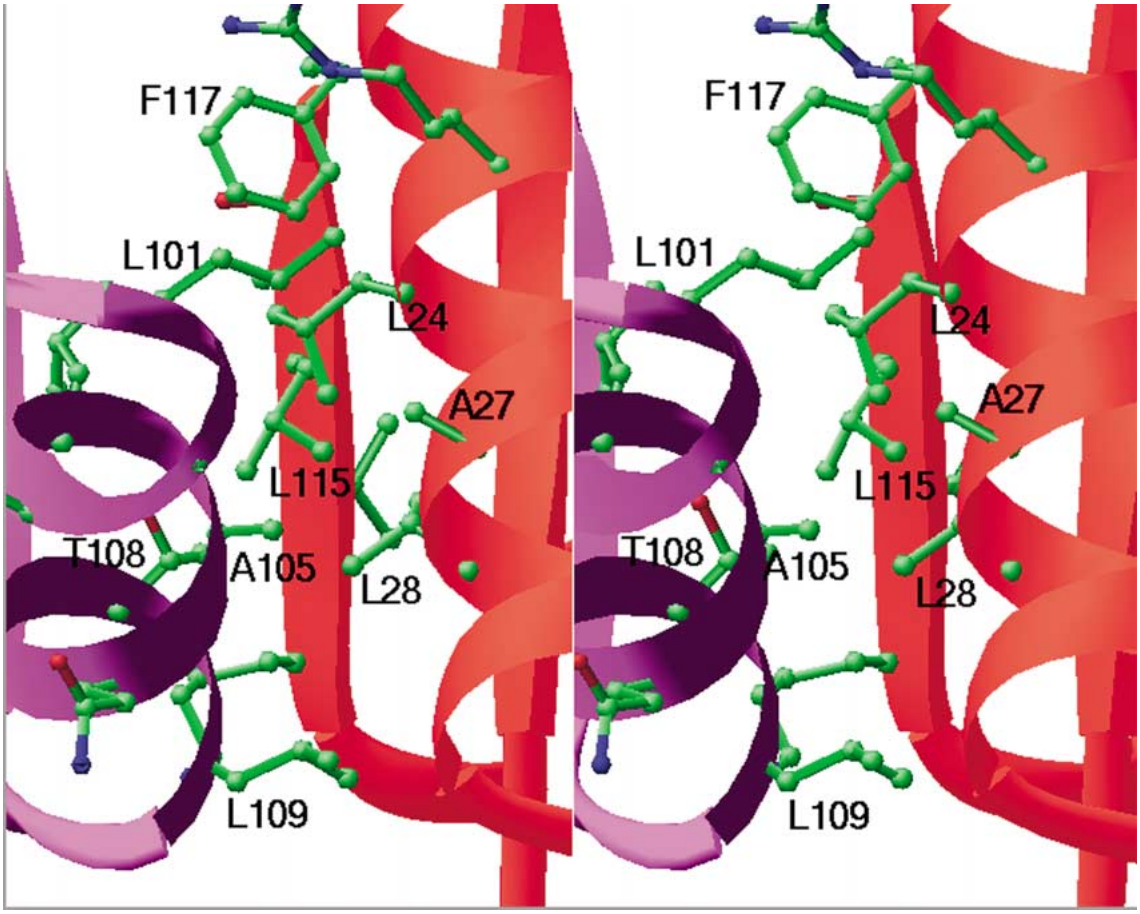

Figure 5a. The hydrophobic interactions between monomers. The hydrophobic residues are labeled and represented by a ball-and-stick model. Others are drawn by a ribbon presentation. The red ribbon represents subunit M1 and the pink ribbon represents subunit M3 as in Figure 4.

4.0 and 11 for z above 5.0). Good structural similarity was observed from 11 homologs despite weak sequence similarity (<18% identity) (Table 3a). The r.m.s deviations are from 1.9 (for 62 pairs of aligned Cα atoms for UDP-N-acetylmuramoyl-L-alanine: D-glutamate ligase) to 3.5 (for 80 pairs of Cα atoms for trimethylamine dehydrogenase). In the SCOP and CATH classification of protein structures, the above 11 structures found in the DALI search can be considered as having the 3-layer (αβα) sandwich architecture similar to the ychN structure. Architecturally, the ychN looks similar to the members of regulatory factor Nef (1superfamily) in SCOP classification or to those of the Rossmann fold in the CATH classification. However, as shown in Table 3a, the topology of the central β-sheet of ten homologs is $3_{\uparrow}2_{\uparrow}1_{\uparrow}4_{\uparrow}5_{\uparrow}$ unlike the topology $5_{\uparrow}1_{\uparrow}2_{\uparrow}3_{\uparrow}4_{\uparrow}$ found in ychN.

There are seven single domain structures with less than 160 amino acid residues among the candidates having a z score above 3.0 in the DALI search. These are listed in Table 3b. Their structures look similar to that of ychN. However, the central β-sheet topology and overall folds are still different from those of ychN.

We also searched the database using a topological relationship supplied by the European Bioinformatics Institute (Web site: *http://www3.ebi.ac.uk/tops/*). The best hits belong to domains from the following enzymes; glutaminase-asparaginase from *Acinetobactor glutaminasificans* (1agx), L-asparaginase from *Wolinella succinogenes* (1wsa), L-asparaginase from *E. coli* (3eca) and glutaminase-asparaginase from *Pseudomonas 7A* (3pga and 4pga). All of these domains belong to 3.40.50.40 in CATH classification and its architecture belongs to a 3-layer (αβα) sandwich and topology in a Rossmann fold. However, none of these central β-sheets have topology $5_{\uparrow}1_{\uparrow}2_{\uparrow}3_{\uparrow}4_{\uparrow}$. Besides, this family shares the same catalytic mechanism

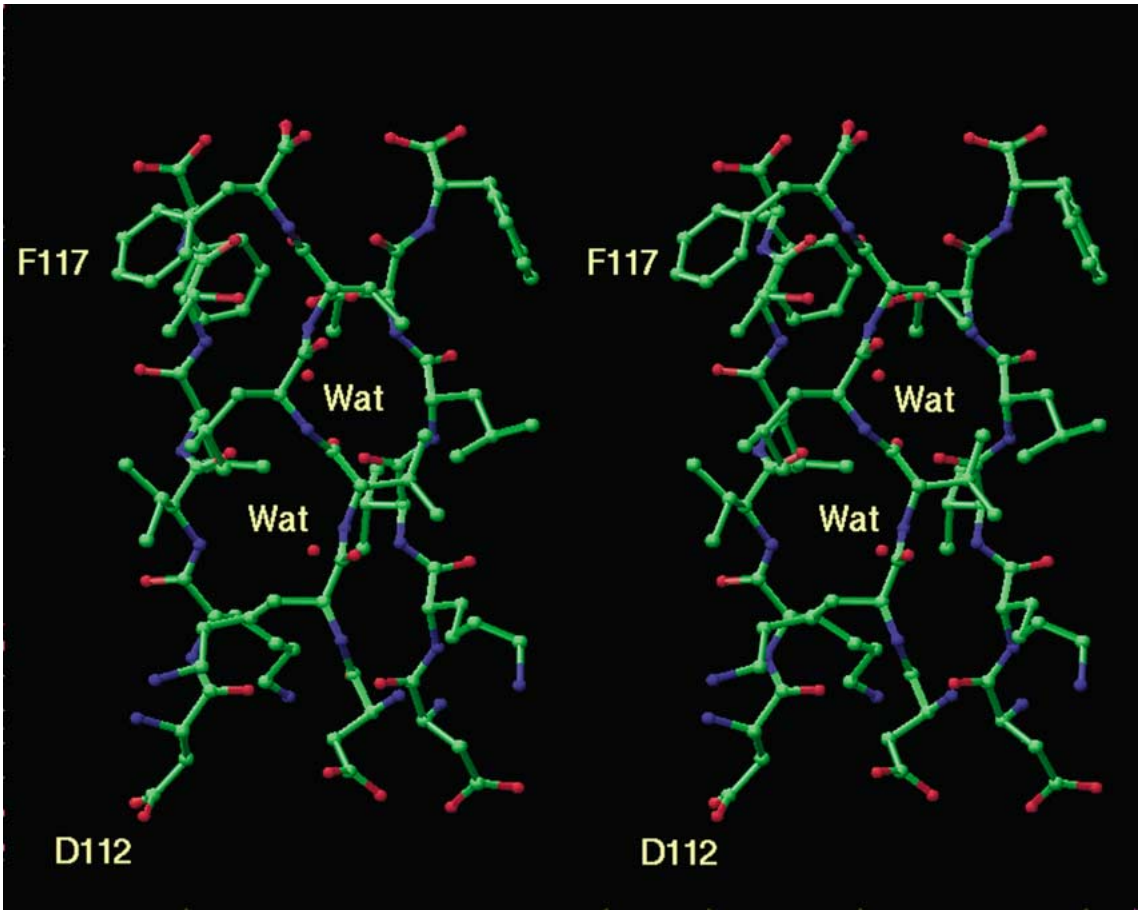

Figure 5b. The interactions among the β-strands (β5s) in the center of the trimer. The residues from Asp112 to Phe117 are represented by a ball-and-stick model. Two water molecules are located in the shallow tunnel formed by β5s. Blue represents nitrogen atoms, red for oxygen, and green for carbon.

involving a characterizing catalytic triad, TKD [29], which are not found in the ychN structure (Figure 8).

In summary, structural homology search with CATH, DALI/FSSP, and SCOP indicates that this protein has a new fold with no obvious similarity to those of other proteins with known three-dimensional structures.

Database search for possible molecular functions

The ychN structure revealed a deep cleft or a pocket, a common feature of an active site. In order to identify the possible active site in ychN, two structural databases containing known structures and functions were queried for a similar residue constellation in ychN. None of the active site templates in the PROCAT database [30] of functional groups in enzyme active sites matches any constellation of residues in the ychN structure. However, a database search for the presence of a known protein motif by RIGOR [31] gave 14 motifs that are found in the ychN structure. Six of them match the clusters of hydrophobic residues of known protein structures. Other motifs are related with the binding of substrates or metal ions, such as in bacteriopheophytin A (1aij; Phe117, Ala8, Leu22, Leu65), dimethylsulfoxide (1cxs; Tyr59, Tyr13, Ser15), thiamine diphosphate (1bfd; Gly14, Ser15, Ala45), magnesium ions (1bfd2; Asn20, Leu22, Arg23), sulfate ion (1fkk; Thr47, Ala48, Gly49 or 1gtm; Ala8, Asn9, Gly10), α-D-mannose (1gai; Ala8, Gly10, Ser43), or acetate ion (1xva; Ala25, Ala27, Glu30). However, we found no indication of bound substrates or metal ions in the electron density maps. The residues predicted to be involved in binding are not highly conserved in the ychN family.

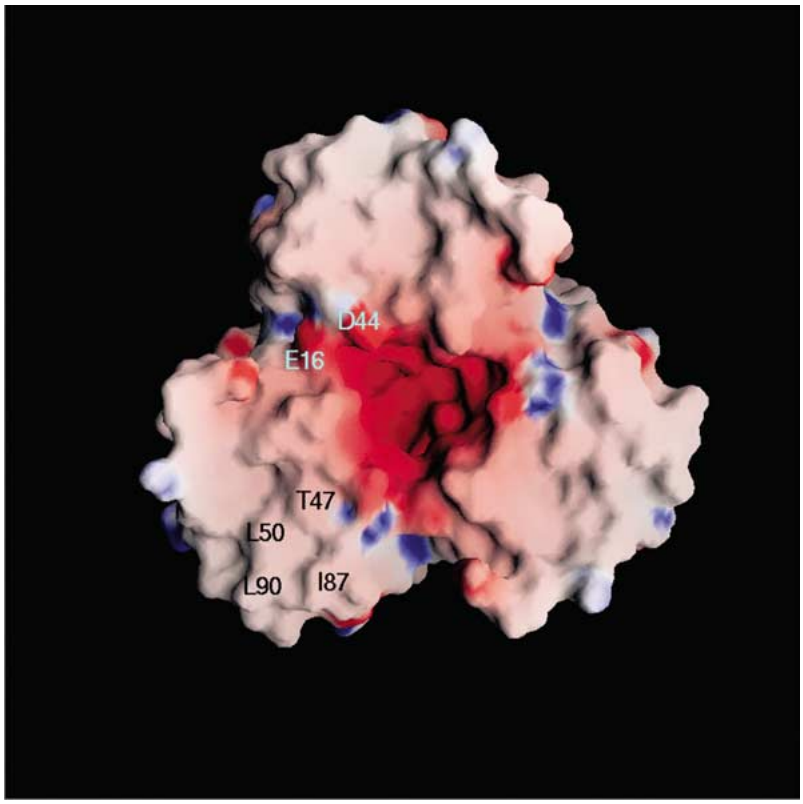

Figure 6. The electrostatic surface potential of the ychN trimer. A molecular surface is drawn along the triad axis cross section created by the program GRASP (red, negative; blue, positive; white, uncharged) [40]. The electrostatic potential and the molecular surface were calculated using the protein atoms. The residues involved in constructing a hydrophobic patch are labeled. Two highly conserved residues, Glu16 and Asp44, lining the wall of the closed cavity are labeled. In this figure, Asp44 belongs to a different subunit from the rest of the labeled residues.

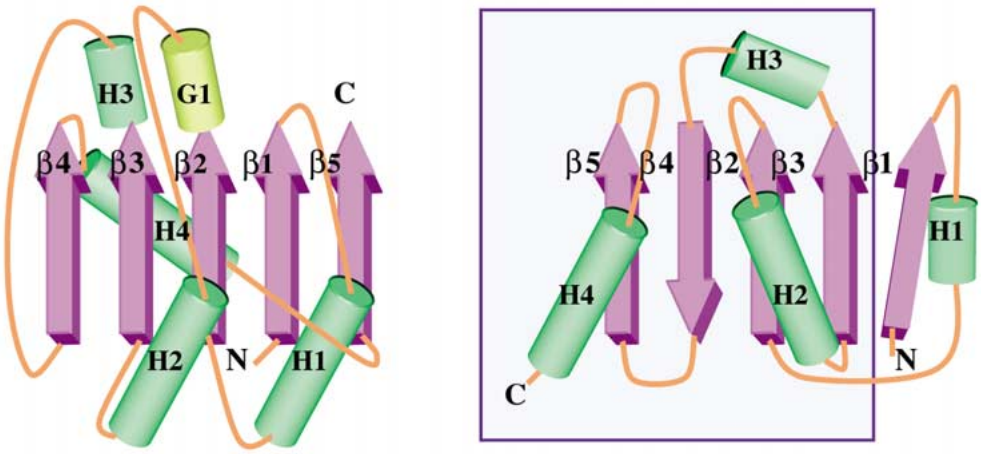

Figure 7. Topology diagrams of ychN and thioredoxin from *E. coli*. α-helices are represented by light green cylinders, β-strands by pink thick arrows, and 3_{10}-helix by a gold cylinder. Secondary structure elements of helices and strands are labeled. In Figure 7b, the blue boxed region represents the typical thioredoxin fold.

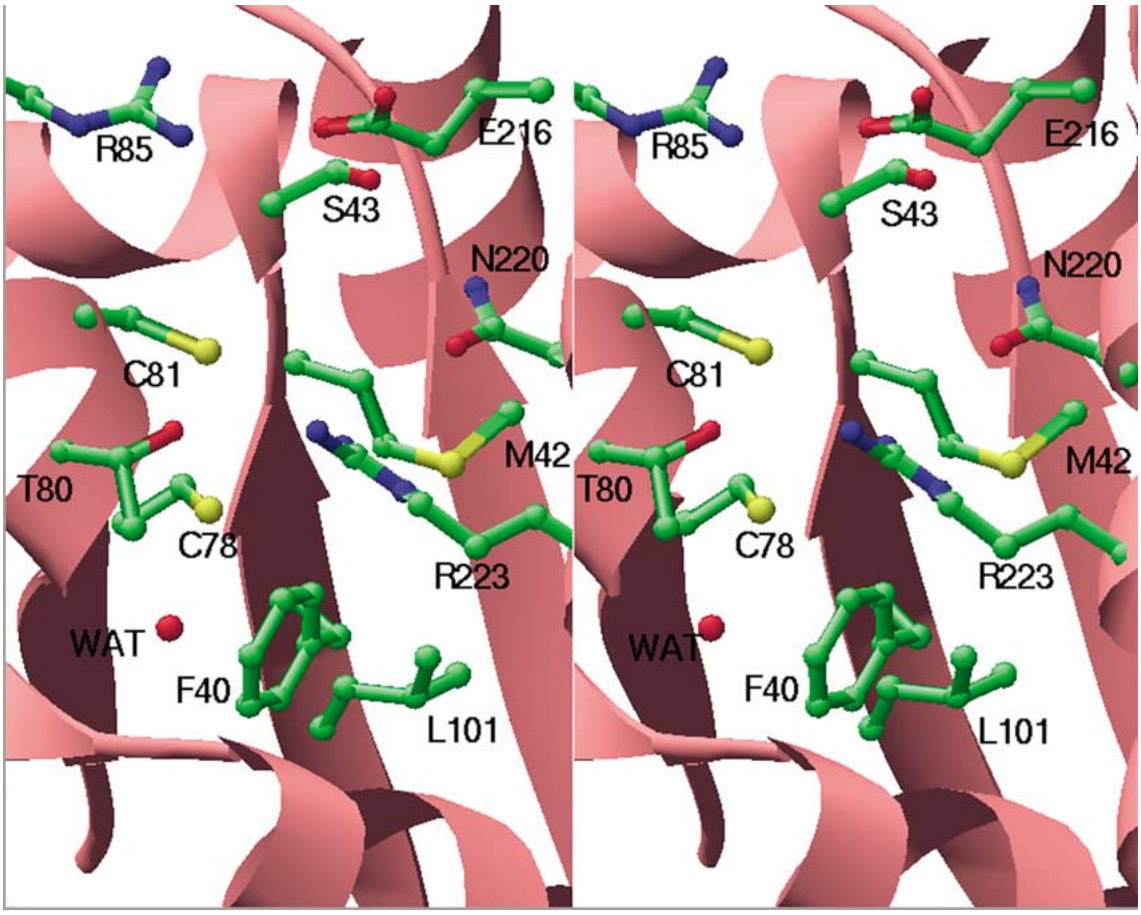

Figure 8. The environment of a possible active site. The residues around two cysteines are labeled and represented by a ball-and-stick model. Others are drawn by a ribbon presentation.

Possible active site and functional implications

Near the equatorial surface of the hexamer, the six interfaces among the adjoining subunits contain recessed cavities (Figure 4). Each cavity is formed by a floor and four walls. The cavity has a depth of ∼11 Å and a width of about 6 Å × 14 Å. The highly conserved residues, Glu16, Arg23, Cys78, Cys81 and Arg85, in the cavity suggested this region as a possible active site (Figure 8). The hydrophobic environment around the active cysteine is frequently observed in many redox proteins [32]. In the case of 2-Cys peroxiredoxin, the hydrophobic environment, coupled with the limited solvent access, appears to preserve the thiol status of the active cysteine prior to its involvement in the catalytic cycle [33]. Considering that Cys78 is an invariant residue in the ychN family, the structural environment around Cys78 implies that this residue is a putative active cysteine.

Another possible function of ychN is that of a thiol:disulfide oxidoreductase. Generally, these oxidoreductases share an active site containing two cysteines, arranged in a C-X-X-C motif, which are either in the reduced state forming two thiols or in the oxidized state forming an intramolecular disulfide bond [32]. Despite very low primary sequence similarity, most of these proteins show the same overall tertiary structure known as the thioredoxin-like fold (Figure 7b) [32]. Sequence alignment with seven ychN homologous proteins obtained from a PSI-BLAST search shows a C-X-X-C motif (Figure 1).

Yet another possibility is hydrolase activity as shown in cysteine proteases such as cathepsin or caspase families. The cysteine proteases are divided into at least 21 families on the basis of sequence or tertiary structures [34]. From 292 candidates obtained from a DALI search, only one belongs to a cysteine protease family. This is a human *Caenorhabditis ele-*

Table 3a. Protein structures similar to ychN, according to DALI.

Name and description	PDB code	Z-score[a]	RMSD[b]	LALI[c]	LESQ[d]	Topology[e]
Flavin oxidoreductase from *E. coli*	1qfj-A	5.7	3.1	86	226	$3_{\uparrow}2_{\uparrow}1_{\uparrow}4_{\uparrow}5_{\uparrow}$
3-phosphoglycerate kinase from *Trypanosoma brucei*	16pk	5.6	3.2	89	415	$3_{\uparrow}2_{\uparrow}1_{\uparrow}4_{\uparrow}5_{\uparrow}$
Trimethylamine dehydrogenase from *Methylophilus methylotrophus* W3A1	2tmd-A	5.3	3.5	80	729	$3_{\uparrow}2_{\uparrow}1_{\uparrow}4_{\uparrow}5_{\uparrow}$
Methionyl-tRNAfmet formyltransferase from *E. coli*	1fmt-A	5.3	2.5	68	308	$3_{\uparrow}2_{\uparrow}1_{\uparrow}4_{\uparrow}5_{\uparrow}6_{\downarrow}7_{\uparrow}$
UDP-N-acetylmuramoyl-L-alanine: D-glutamate ligase from *E. coli*	4uag-A	5.1	1.9	62	429	$3_{\uparrow}2_{\uparrow}1_{\uparrow}4_{\uparrow}5_{\uparrow}$
Malate dehydrogenase from *E. coli*	2cmd	5.1	2.5	67	312	$3_{\uparrow}2_{\uparrow}1_{\uparrow}4_{\uparrow}5_{\uparrow}6_{\uparrow}7_{\downarrow}8_{\uparrow}$
Carbonyl reductase from mouse lung	1cyd-A	5.1	2.8	70	242	$3_{\uparrow}2_{\uparrow}1_{\uparrow}4_{\uparrow}5_{\uparrow}6_{\uparrow}7_{\uparrow}$
Spoiiaa from *Bacillus subtilis*	1auz	5.1	3.3	87	116	$1_{\downarrow}2_{\uparrow}3_{\uparrow}4_{\uparrow}5_{\uparrow}$
Alcohol dehydrogenase from *Drosophila lebanonensis*	1b16-A	5.0	2.8	69	254	$3_{\uparrow}2_{\uparrow}1_{\uparrow}4_{\uparrow}5_{\uparrow}6_{\uparrow}7_{\uparrow}8_{\downarrow}$
D-amino acid oxidase from *Sus Scrofa*	1an9-A	5.0	3.4	79	340	$3_{\uparrow}2_{\uparrow}1_{\uparrow}4_{\uparrow}6_{\uparrow}5_{\downarrow}$

[a]Z-score, strength of structural similarity in standard deviations. Only matches above a threshold of $Z = 5$ are reported.
[b]RMSD, positional root mean square deviation of superimposed Cα atoms in Angstroms.
[c]LALI, total number of equivalenced residues.
[d]LSEQ2, length of the entire chain of the equivalenced structure.
[e]Topology, topology of a central β-sheet in the equivalenced region.

Table 3b. Single domain structures similar to ychN, according to DALI.

Name and description	PDB code	z score	RMSD	LALI	LESQ	Topology
Spoiiaa from *Bacillus subtilis*	1auz	5.1	3.3	87	116	$1_{\downarrow}2_{\uparrow}3_{\uparrow}4_{\uparrow}5_{\uparrow}$
ATP-binding domain of protein Mj0577 from *Methanococcus jannaschii*	1mjh-A	4.3	3.3	78	143	$3_{\uparrow}2_{\uparrow}1_{\uparrow}4_{\uparrow}5_{\uparrow}$
3-dehydroquinate dehydratase from *Salmonella typhi*	2dhq-A	4.2	3.5	84	136	$2_{\uparrow}1_{\uparrow}3_{\uparrow}4_{\uparrow}5_{\uparrow}$
CheY from *Thermotoga maritima*	1tmy	3.7	3.1	75	118	$2_{\uparrow}1_{\uparrow}3_{\uparrow}4_{\uparrow}5_{\uparrow}$
Transcriptional regulatory protein Fixj from *Sinorhizobium meliloti*	1dbw-A	3.7	3.4	76	123	$2_{\uparrow}1_{\uparrow}3_{\uparrow}4_{\uparrow}5_{\uparrow}$
CheY from *E. coli*	3chy	3.3	2.3	47	128	$2_{\uparrow}1_{\uparrow}3_{\uparrow}4_{\uparrow}5_{\uparrow}$
Lumazine synthase from *Brucella abortus*	1di0-A	3.3	3.9	80	148	$2_{\uparrow}1_{\uparrow}3_{\uparrow}4_{\uparrow}$
Phosphopantetheine adenylyltransferase from *E. coli*	1b6t-A	3.2	3.1	71	157	$3_{\uparrow}2_{\uparrow}1_{\uparrow}4_{\uparrow}5_{\uparrow}$

gans (CED-3) homologue, apopain/CPP32 (1pau, z = 2.4) which is a cysteine protease related to mammalian interleukin-1β converting enzyme (ICE) [35]. However, as the ychN structure does not possess a histidine residue at the active site, we would have to propose a novel mechanism involving a different residue if it is indeed a cysteine protease.

Thermostability of ychN

Although ychN is an *E. coli* protein, it was stable at 80 °C during purification. The crystal structure gives possible reasons for the unusual thermostability of this protein. Salt bridges are one of the major factors for thermostability. Hyperthermophilic enzymes in general possess a much higher number of ion pairs per residue. This number is 0.04 in normal enzymes, but is 0.085 in the hyperthermophilic tungstoprotein enzyme, aldehyde ferredoxin oxidoreductase (AOR) from *Pyrococcus furious* [36]. The *E. coli* ychN structure reveals seven salt bridges, and 0.060 ion pairs per residue in both the monomer and hexamer. As this exceeds the average value, salt bridges may be important in the thermostability of ychN. One of salt bridges, between Arg85 with Glu16 or Asp44, forms a tertiary salt bridge known to increase thermostability by stabilizing helix dipole [37].

Another factor is a reduced surface area and optimized packing of the atom in the core of the structure. In case of the extremely thermostable AOR, the accessible surface is reduced substantially upon oligomerization [36] as indicated by the low ratio of Ao/Ac, where Ao and Ac are the observed and calculated surface area, respectively. These values are 0.94 and 0.84 for monomeric and hexameric ychN, respectively. A simple indicator for evaluating the efficiency of packing is given by the fraction of atoms in a protein with zero accessible surface area. For the hyperthermophilic AOR, the fraction (∼0.55) is significantly higher than the average (∼0.5) [36]. For the monomeric ychN, this fraction is 0.53. Furthermore, this number increases to 0.60 in the hexameric form. These indicate that ychN itself has a more tightly packed structure than average proteins and hexamerization makes the whole structure even more compact. Therefore, a tightly packed structure and presence of salt bridges can be a plausible explanation for the thermostable nature of ychN.

Acknowledgements

We thank Dr Thomas Earnest and Dr Keith Henderson (Advanced Light Source, Lawrence Berkeley National Laboratory) for assistance during data collection at ALS, Dr Weiru Wang and Dr Ed Berry for helpful advice during the structure determination, and Dr Igor Grigoriev for searching the *E. coli* database. This work was supported by the grants from the Director, Office of Science, Office of Biological and Environmental Research under U.S. Department of Energy (Contract No. DE-AC03-76SF00098) and from National Institutes of Health (P50 GM62412).

References

1. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) *Nucleic Acids Res.* **25**, 3389–3402.
2. Murzin, A. G. and Patty, L. (1999) *Curr. Opin. Struct. Biol.* **9**, 359–361.
3. Marcotte, E. M., Pellegrini, M.,Thompson, M. J., Yeates, T. O., and Eisenberg, D. (1999) *Nature* **402**, 83–86.
4. Zarembinski, T. I., Hung, L.-W., Mueller-Dieckmann, H.-J., Kim, K.-K., Yokota, H., Kim, R. and Kim, S.-H. (1998) *Proc. Natl. Acad. Sci. USA.* **95**, 15189–15193.
5. Hwang, K. Y., Chung, J. H., Kim, S.-H, Han, Y. S., and Cho, Y. (1999) *Nat. Struct. Biol.* **6**, 691–696.
6. Teplova, M., Tereshko, V., Sanishvili, R., Joachimiak, A., Bushueva, T., Anderson, W. F. and Egli, M. (2000) *Protein Sci.* **9**, 2557–2566.
7. Yang, F., Gustafson, K. R., Boyd, M. R., and Wlodawer, A. (1998*) Nat. Struct. Biol.* **5**, 763–764.
8. Colovos, C., Cascio, D. and Yeates, T. O. (1998) *Structure Fold Des.* **6**, 1329–1337.
9. Blattner, F. R., Plunkett, G. I. I. I., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B. and Shao,Y. (1997) *Science* **277**, 1453–1474.
10. Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D. and Koonin, E. V. (2001) *Nucleic Acids Res.* **29**, 22–28.
11. Pott, A. S. and Dahl, C. (1998) *Microbiology* **144**, 1881–1894.
12. Leahy, D. J., Hendrickson, W. A., Aukhil, I. and Erickson, H. P. (1992) *Science* **258**, 987–991.
13. Kim, R., Sandler, S. J., Goldman, S., Yokota, H., Clark, A. J. and Kim, S.-H. (1998) *Biotech. Lett.* **20**, 207–210.
14. Jancarik, J. and Kim, S.-H. (1991) *J. Appl. Crystallogr.* **24**, 409–411.
15. Otwinowski, Z. and Minor, W. (1997) *Methods Enzymol.* **276**, 307–326.
16. Terwilliger, T. C. and Berendzen, J. (1999) *Acta Crystallogr. D* **55**, 849–861.
17. Dodson, E. J., Winn, M. and Ralph, A. (1997) *Methods Enzymol.* **277**, 620–633.
18. Jones, A. and Kleywegt, G. (1997) *Methods Enzymol.* **277**, 173–208.

19. Kleywegt, G. J. and Jones, T. A. (1999) *Acta Crystallogr. D* **55**, 941–944.
20. Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. and Warren, G. L. (1998) *Acta Crystallogr. D* **54**, 905–921.
21. Luzzati, V. (1952) *Acta Crystallogr.* **5**, 802–810.
22. Laskowski, R. A., MacArthur, M. W., Moss, D. S. and Thornton, J. M. (1993) *Appl. Crystallogr.* **26**, 283–291.
23. Dasgupta, S., Iyer, G. H., Bryant, S. H., Lawrence, C. E. and Bell, J. A. (1997) *Proteins* **28**, 494–514.
24. Richardson, J. S. (1981) *Adv. Protein Chem.* **34**, 167–339.
25. Zhang, C. and Kim, S.-H. (2000) *J. Mol. Biol.* **299**, 1075–1089.
26. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. and Thornton, J. M. (1997) *Structure Fold Des.* **5**, 1093–1108.
27. Holm, L. and Sander, C. (1997) *Nucleic Acids Res.* **25**, 231–234.
28. Hubbard, T. J., Murzin, A. G., Brenner, S. E. and Chothia, C. (1997) *Nucleic Acids Res.* **25**, 236–239.
29. Lubkowski, J., Palm, G. J., Gilliland, G. L., Derst, C., Rohm, K. H., and Wlodawer, A. (1996) *Eur. J. Biochem.* **241**, 201–207.
30. Wallace, A. C., Borkakoti, N. and Thornton, J. M. (1997) *Protein Sci.* **6**, 2308–2323.
31. Kleywegt, G. J. (1999) *J. Mol. Biol.* **285**, 1887–1897.
32. Fabianek, R. A., Hennecke, H. and Thony-Meyer, L. (2000) *FEMS Microbiol Rev.* **24**, 303–316.
33. Schroder, E., Littlechild, J. A., Lebedev, A. A., Errington, N., Vagin, A. A. and Isupov, M. N. (2000) *Structure Fold Des.* **8**, 605–615.
34. Otto, H.-H. and Schirmeister, T. (1997) *Chem. Rev.* **97**, 133–172.
35. Rotonda, J., Nicholson, D. W., Fazil, K. M., Gallant, M., Gareau, Y., Labelle, M., Peterson, E. P., Rasper, D. M., Ruel, R., Vaillancourt, J. P., Thornberry, N. A. and Becker, J. W. (1996) *Nat. Struct. Biol.* **3**, 619–625.
36. Chan, M. K., Mukund, S., Kletzin, A., Adams, M. W. W. and Rees, D. C. (1995) *Science* **267**, 1463–1469.
37. Das, R. and Gerstein, M. (2000) *Funct. Integr. Genomics* **1**, 76–88.
38. Carson, M. (1991) *J. Appl. Crystallogr.* **24**, 958–961.
39. Kraulis, P. J. (1991) *J. Appl. Crystallogr.* **24**, 946–950.
40. Nicholls, A. (1991) *Proteins* **11**, 281–296.